COMBINING DEEP AND SHALLOW NEURAL NETWORKS WITH AD HOC DETECTORS FOR THE CLASSIFICATION OF COMPLEX MULTI-MODAL URBAN SCENES

Daniele Cerra, Miguel Pato, Emiliano Carmona, Seyed Majid Azimi, Jiaojiao Tian, Reza Bahmanyar, Franz Kurz, Eleonora Vig, Ksenia Bittner, Corentin Henry, Pablo d'Angelo, Rupert Müller, Kevin Alonso, Peter Fischer, and Peter Reinartz

Remote Sensing Technology Institute (MF-PBA), German Aerospace Center (DLR),
Münchnerstr. 20, 82234 Weßling, Germany

ABSTRACT

This article describes the workflow of the classification algorithm which ranked at 2^{nd} place in the 2018 GRSS Data Fusion Contest. The objective of the contest was to provide a classification map with 20 classes on a complex urban scenario. The available multi-modal data were acquired from hyperspectral, LiDAR and very high-resolution RGB sensors flown on the same platform over the city of Houston, TX, USA. The classification was obtained by merging deep convolutional and shallow fully-connected neural networks on a simplified set of classes, complemented by a series of specific detectors and ad hoc classifiers.

Index Terms— data fusion, classification, LiDAR, hyperspectral, very high-resolution

1. INTRODUCTION

The multi-modal dataset distributed in support of the GRSS Data Fusion Contest 2018[1] comprised data acquired from hyperspectral, LiDAR and very high-resolution RGB sensors, which were flown on the same platform over the city of Houston, TX, USA. The complementary information yielded from these datasets needs to be exploited in order to assign each pixel to one of the 20 defined semantic classes (for a list of the classes see Table 1). The semantic labelling of urban areas often involves classes of interest differing greatly in spectral, textural and other higher-order features: spectral features may drive the detection of healthy grass, height information is vital for classes such as buildings, and shape features aid in the detection of vehicles. Therefore, it is difficult to find a single classifier which correctly identifies all the different classes in such tasks.

[1]The authors would like to thank the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory at the University of Houston for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. Data available at http://www.grss-ieee.org/community/technical-committees/data-fusion/data-fusion-contest

Recently, classifiers based on deep learning have proven very promising in capturing the relevant features from a wide variety of classes; however, these may over-rely on higher order interactions among the pixels composing an object. In this article, we use a deep convolutional neural network (CNN) together with a shallow fully-connected neural network (NN). Natural classes driven by spectral and simple textural properties (e.g., grass, trees, water) are extrapolated from the NN classification and overlaid on the output of the CNN in which they are under-represented and yield a considerable number of false negatives. The classification is initially carried out on a simplified set of classes and completed by using (1) ad hoc spectral detectors to improve bare soil characterization, (2) separately trained neural networks to discriminate different kinds of buildings and to detect vehicles, and (3) template matching to find crosswalks.

The final classification results yielded an overall accuracy (OA) of 80.74% and ranked second in the contest, with negligible distance (0.04%) from the best classification.

2. PREPROCESSING AND FEATURE EXTRACTION

The dataset provided by the contest included multispectral (MS) LiDAR at 50 cm ground sampling distance (GSD), hyperspectral (HS) at 1 m GSD, and very high-resolution RGB imagery at 5 cm GSD. To prepare the dataset for classification, a number of preprocessing steps were carried out. Normalized digital surface models (nDSMs) were generated by subtracting a low-passed digital terrestrial model from the LiDAR digital surface models (first and last pulses) and removing additional noise. The MS LiDAR intensity images were denoised by a 5×5 median filter, while the RGB images were mosaicked and down-sampled to 50 cm GSD. From the HS image, 42 bands were selected and up-sampled to 50 cm GSD using an order-3 spline. For the generic base classification (see Section 3), an input stack is generated at 50 cm GSD in which each pixel is represented by a 100-D vector, comprising 42 HS bands, 3 RGB bands, 3 MS LiDAR intensity bands,

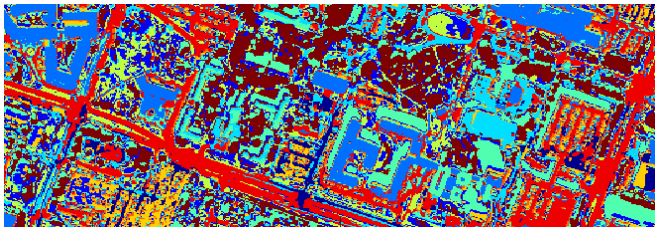


Fig. 1. Color illustration of the topics. For visualization, each pixel is assigned to the largest-value topic in its topic vector.

2 nDSMs, and a 50-D topic vector.

The topic vectors are multi-modal high-level features computed by applying multi-modal latent Dirichlet allocation (mmLDA) [1] to the bag-of-words (BoW) model of the HS (1 m GSD) and RGB (50 cm GSD) images. The mmLDA discovers the joint model latency as a set of so-called topics and represents each image as topic mixtures. To compute these features, the RGB and HS images are tiled into patches of 32×32 and 16×16 pixels, respectively. Then their local primitive features are extracted by vectorizing a window of 3×3 pixels around each pixel, resulting in 9-D feature vectors. Each image band is treated separately and the feature vectors are concatenated afterwards, resulting in 27-D and 378-D feature vectors for the RGB and HS images, respectively. The image patches are then modeled as BoWs, and mmLDA is employed to discover their joint latency as 25 topics, with each image pixel represented as a mixture of the topics (topic vector). Finally, for each pixel the HS (up-sampled to 50 cm GSD) and RGB topic vectors are concatenated forming a 50-D vector. Fig. 1 shows a color illustration of the topics.

For the specific detectors, additional features were extracted from the HS images, namely the normalized difference vegetation index (NDVI) computed by using the HS bands 28 for near infrared (789 nm) and 18 for red (646 nm), and 13 minimum noise fraction (MNF) components.

3. CLASSIFICATION

Our approach to the classification task was to combine generic base classifiers and a series of specific detectors. The task of the base classifiers was simplified by merging classes 1-2 (grass) and 8-9 (buildings), while ignoring the classes 12 (crosswalks) and 18 (cars). We implemented two base classifiers for the remaining 16 classes: a deep CNN and a shallow fully-connected NN. Both networks were implemented and trained using the Keras API[2] with the TensorFlow backend.

CNN - The input to the CNN consisted of cubes of 5×5 pixels $\times$ 50 features (HS, RGB, MS LiDAR, nDSMs as discussed in Section 2), which were reshaped to 2-D matrices of 25 spatial pixels $\times$ 50 spectral features. The network consisted of 8 convolution layers applied selectively along the spatial or spectral direction, or both. The same 1-D convolution along the spatial dimension was applied to the stack of HS, RGB, MS LiDAR, and nDSM features. Different feature groups were then again separated and went through 2-D convolutional layers. Finally, all features were stacked together in a final layer, where the topics vector could be optionally incorporated. A dense hidden layer was used to connect the network to a softmax layer, employing categorical cross entropy as loss function. For optimization, the Adam optimizer was used with the *amsgrad* option. In order to reduce overfitting, L2 regularization was applied to all convolutional layers and a 25% dropout was used after the last fully connected layer. The training stopped after a small number of epochs (2-6).

Shallow NN - The NN base classifier consisted of a fully-connected neural network with two hidden layers (128 and 64 nodes, ReLU activations), a final softmax layer, and a categorical cross entropy loss function. The full stack of 100-D features including HS, RGB, MS LiDAR, nDSMs, and topic vector was used as input. The network was trained for 5 epochs employing stochastic gradient descent with a batch size of 128 and applying class weights inversely proportional to the class frequencies.

CNN vs. NN - The CNN performed better than the NN classifier on classes strongly characterized by context, such as different types of roads. However, natural classes such as grass, trees and water (strongly related to their spectral properties) yielded numerous false negatives upon visual inspection. Taking advantage of the fact that such classes yielded no false positives in CNN, it was possible to overlay the output of the NN classifier for these classes only. Single pixels for a class overlaid from the NN classifier could therefore belong to the CNN classification, while the opposite cannot happen. This allowed avoiding to explicitly perform data fusion at decision level. The classifier of choice for each class is reported in Table 1, along with the final OA for the class.

3.1. Ad Hoc Classifiers

After the base classification, the following processing steps were carried out and the obtained maps were overlaid in the same order to derive the final classification:

Bare Earth - An additional soil map for the class 'bare earth' was derived from the HS image using spectral angle mapper (SAM) applied on the same set of simplified classes used in input for the CNN and NN classifiers. Results were regularized by two morphological openings and closings using a disk-shape structuring element with radius 2.

Commercial and Residential Buildings - The residential and commercial buildings were first detected by thresholding (height $>$ 1 m) the normalized last pulse LiDAR DSM. Morphological openings and closings refined the detected build-

[2]https://keras.io

Index		Class	Classifier	OA (%)
1		Healthy grass	NN	94.5
2		Stressed grass		88.7
3		Artificial turf	NN	95.7
4		Evergreen trees	NN	96.5
5		Deciduous trees	NN	81.6
6		Bare earth	NN	94.0
7		Water	NN	90.8
8		Residential buildings	CNN	83.1
9		Commercial buildings		90.6
10		Roads	CNN	70.4
11		Sidewalks	CNN	60.3
12		Crosswalks	None	30.6
13		Major thoroughfares	CNN	35.7
14		Highways	CNN	72.4
15		Railways	CNN	93.2
16		Paved parking lots	NN	65.6
17		Unpaved parking lots	CNN	0.0
18		Cars	None	97.0
19		Trains	CNN	93.4
20		Stadium Seats	NN	92.4

Table 1. Selected classifier and OA per class. Classes denoted by "none" were not considered in the initial CNN and NN classifications. Colors represent a legend for Fig. 3(c).

ing masks, which were then separated into residential and commercial by using a region-based random forests classifier applied to the spectral, height, and shape features according to [2]. The classification results were refined based on the features automatically learned by an end-to-end fully convolutional neural network (FCN), consisting of two parallel networks merged at a late stage to integrate the spectral and height information from RGB and nDSM data.

Stadium Seats - An ad hoc binary NN classifier (with the same architecture as the base NN classifier) was trained to detect stadium seats against all other classes. The input included the MNF components derived from the HS image, and the nadir and DSM bands provided with HS data.

Healthy and Stressed Grass - The stressed and healthy grass (treated as a single class during the base classification) were separated by hard thresholding the NDVI at 0.535.

Highway - The highway was regularized by exploiting its low curvature. The Hough transform was applied to the binary map of the class to derive the three most dominant linear features. All pixels within a certain distance from these lines (based on the highway width) were then assigned to the highway class if they had been originally classified as roads or major thoroughfares.

Morphological Filtering - The results were regularized by three cycles of morphological opening and closing using a disk-shape structuring element with radius 2.

Crosswalks - A normalized cross correlation matching algorithm was trained on three of the crosswalks contained in the ground truth and applied to the RGB mosaic at 5 cm GSD. The sizes of the trained samples were slightly changed during matching to reduce the number of false negatives and the output was down-sampled to 50 cm GSD.

Cars - In order to detect cars, a pixel-wise vehicle segmentation algorithm based on FCNs (similar to [3]) was applied to the RGB mosaic at 5 cm GSD. The network was pretrained on DLR's 3K images with 13 cm GSD. Because of the small number of cars annotated in the training dataset, the first pulse of the LiDAR DSM resampled at 5 cm GSD was thresholded at 20 cm to run a semi-automatic annotation in the training dataset. The network was then fine-tuned on the training data and applied to the test dataset. The resulting car mask was down-sampled to 50 cm GSD and refined by performing a morphological opening and dilation using a disc-shape structuring element with radius 1. Fig. 2 shows some detected cars. Finally, the cars detected on highways were discarded as they were yielding false positives.

Fig. 2. Pixel-wise car segmentation.

Fig. 3 presents our final classification map.

4. CONCLUSIONS

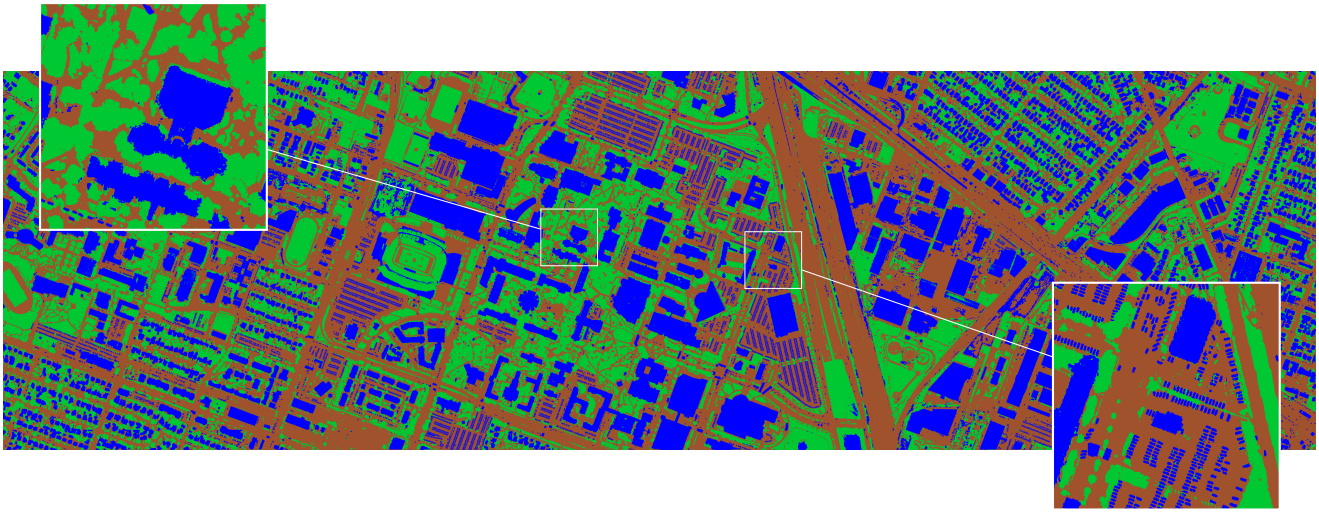
In a complex urban scenario, the classes of interest can differ greatly in shape, scale, spectral features, and complex higher-order statistics. Therefore, it is not always possible to use a single classifier for semantic labelling. In the framework of the GRSS Data Fusion Contest, we employed both deep and shallow neural networks on a simplified set of classes, where classes driven by their spectral properties were obtained from the shallow classifier. Finally, a set of ad hoc detectors based on the different properties and characteristics of each class of interest were used to finalize the classification. The results ranked at second place in the contest with an overall accuracy of 80.74 % and an average accuracy of 76.32 %.

5. REFERENCES

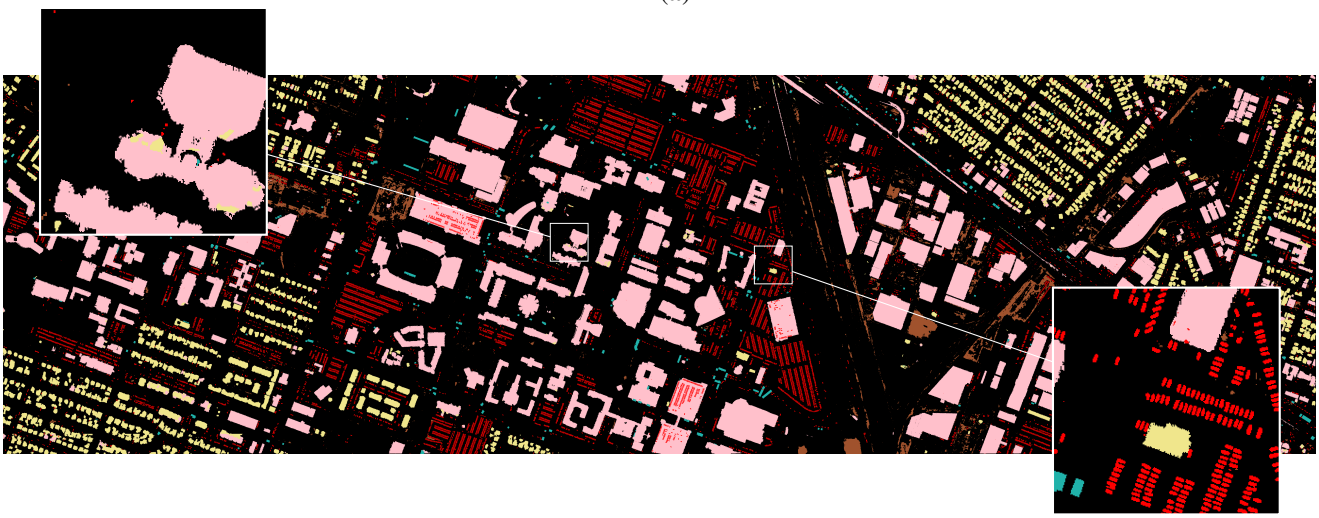
[1] R. Bahmanyar, D. Espinoza-Molina, and M. Datcu, "Multisensor Earth observation image classification based on a multimodal latent Dirichlet allocation model," *IEEE GRSL*, vol. 15, no. 3, pp. 459–463, March 2018.

[2] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE TGRS*, vol. 52, no. 1, pp. 406–417, January 2014.

[3] S. Majid Azimi, P. Fischer, M. Körner, and P. Reinartz, "Aerial LaneNet: Lane Marking Semantic Segmentation in Aerial Imagery using Wavelet-Enhanced Cost-sensitive Symmetric Fully Convolutional Neural Networks," *ArXiv e-prints*, March 2018.



(**a**)



(**b**)

(**c**)

Fig. 3. (**a**) Contribution to the final classification results of deep CNN (sienna), shallow NN classification (green), and ad hoc detectors and classifiers (blue); (**b**) Classes belonging to the ad hoc detectors and classifiers: bare soil (sienna), residential buildings (yellow), commercial buildings (pink), crosswalks (cyan), cars (red). (**c**) Final classification results.